Computer Science and Application 计算机科学与应用, 2021, 11(11), 2731-2741
Published Online November 2021 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2021.1111277

Hans汉斯

基于知识增强卷积神经网络的标的物命名实体识别方法

高振祥[1*]，江　静[1]，陈　建[1#]，刘金硕[2]

[1]国家能源集团，北京
[2]武汉大学国家网络安全学院，湖北 武汉

收稿日期：2021年10月15日；录用日期：2021年11月12日；发布日期：2021年11月22日

摘　要

针对招标文件中，“标的物”作为命名实体存在着分词错误、多个名词并列现象导致的真实意图标的物命名实体提取困难问题，提出一种基于知识增强卷积神经网络(CNN)的标的物命名实体识别方法。该方法首先构建了针对招标文件的正则表达式，实现包含标的物短语的定位。然后利用基于知识增强卷积神经网络，在输入层将标的物定位短语和其上下文信息作为输入，通过卷积层对特征进行提取，最后通过Softmax层输出实体标注结果。在2017~2020年的19,980份招标文件的数据集上，本方法的平均准确率为0.96，与深度神经网络(DNN)、循环神经网络(RNN)和Hopfield神经网络(HNN)相比准确率分别提升了1.2%、0.4%和0.3%。实验结果表明本方法能够进一步提高标的物命名实体识别的准确率，使得企业在智能化标的物提取过程中取得更优的效果。

关键词

标的物命名实体识别，正则表达式，卷积神经网络，自然语言处理，知识增强

Subject Matter Named Entity Recognition Method Based on Knowledge-Enhanced Convolutional Neural Network

Zhenxiang Gao[1*], Jing Jiang[1], Jian Chen[1#], Jinshuo Liu[2]

[1]China Energy Investment, Beijing
[2]School of Cyber Science and Engineering, Wuhan University, Wuhan Hubei

Received: Oct. 15th, 2021; accepted: Nov. 12th, 2021; published: Nov. 22nd, 2021

[*]第一作者。
[#]通讯作者。

文章引用: 高振祥, 江静, 陈建, 刘金硕. 基于知识增强卷积神经网络的标的物命名实体识别方法[J]. 计算机科学与应用, 2021, 11(11): 2731-2741. DOI: 10.12677/csa.2021.1111277

Abstract

In view of the difficulty in extracting the naming entity of the real meaning icon caused by word segmentation error and multiple noun juxtaposition in the bidding document, a subject matter named entity recognition method based on knowledge-enhanced Convolutional Neural Network (CNN) was proposed. Firstly, a regular expression was constructed for the bidding document to locate the phrase containing the subject matter. Then, the knowledge-enhanced convolutional neural network was used to take the target location phrase and its context information as the input in the input layer, extracting the features through the convolution layer, and finally output the entity annotation results through the Softmax layer. On the data set of 19,980 bidding documents from 2017 to 2020, the average accuracy of this method is 0.96, which is improved by 1.2%, 0.4% and 0.3% respectively compared with Deep Neural Network (DNN), Recurrent Neural Network (RNN) and Hopfield Neural Network (HNN). The experimental results show that this method can further improve the accuracy of object named entity recognition, and make enterprises achieve better results in the process of intelligent object extraction.

Keywords

Subject Matter Named Entity Recognition, Regular Expression, Convolutional Neural Network (CNN), Natural Language Processing, Knowledge-Enhanced

1. 引言

招标业务是企业进行项目管理的一项重要工作，而企业对于招标项目管理过程中的标的物提取依然采用人工方式进行处理，消耗了大量不必要的人力物力。因此，可以利用计算机实现标的物自动提取，从而显著提高相关领域企业工作人员的工作效率，有效提高招标的质量，促进企业对招标的管理方式在智能化、电子化的大方向上发展。

特定命名实体识别的主要任务是识别出文本中的人名、地名、化学名词等不能用通用名词构成规则和长度而划分的专有名词[1] [2] [3]。命名实体是命名实体识别的研究主体，一般包括 3 大类(实体类、时间类和数字类)和 7 小类(人名、地名、机构名、时间、日期、货币和百分比)命名实体[4] [5] [6] [7]。由于数量、时间、日期、货币等实体识别通常可以采用模式匹配的方式获得较好的识别效果，相比之下人名、地名、机构名的构成更复杂和常用，因此这几种实体相关研究是近几年的研究热点[8] [9]。

招标文件文本中由于标的物命名实体会涉及到机构名和专有的物品名称，还与行业领域相关，而且还有多个名词同时并列出现的情况，其识别难度很大。标的物命名实体的成词规则不同于通用的名词构词规则，但是标的物却是招标文件分类的重要参数，因此，针对行业领域的招标文件的标的物命名实体识别开展研究具有重要的理论和现实意义。

首先，本文针对招标文件文本，构建相关的规则集和字段词典作为“知识”，利用正则表达式对标的物命名实体所在的文本片段进行定位；然后，利用基于知识增强卷积神经网络(Convolutional Neural Network, CNN)，在定位的文本片段上进一步识别标的物命名实体，解决多个名词并列相邻以及构词规则特殊化问题，同时还利用了深度学习模型能利用上下文语义关系的优化。

本文第 2 节介绍了与研究内容相关的理论知识和技术内容。第 3 节对标的物提取的思路进行了阐述。第 4 节对实现结果进行了分析。最后对研究内容进行了总结。

2. 相关理论及技术

2.1. 正则表达式

正则，顾名思义就是规则，而正则表达式就是从字符串中总结出规则，并用表达式的形式对这种规则进行总结，从而将这种规则表达出来。正则匹配的总体流程是首先对相关语料进行分析，根据研究目的总结需提取字段的规则，然后正则表达式提取相应字段。其具体实现流程：1) 读取字符并查找状态起始位置；2) 查找要匹配字符；3) 判断是否与匹配字符相同，如果匹配成功，就将该字符进行提取，否则读取下一个字符，重复 2，3 过程直到所有匹配字符都提取完全。

2.2. 基于深度学习的命名实体识别

命名实体识别语言处理中的一项基础任务，应用范围非常广泛。命名实体一般指的是文本中具有特定意义或者指代性强的实体，通常包括人名、地名、组织机构名、日期时间、专有名词等。近几年计算机的计算能力逐渐提升，深度神经网络(Deep Neural Networks, DNN)、长短时记忆网络(Long Short-Term Memory, LSTM)、循环神经网络(Recurrent Neural Networks, RNN)、CNN 等更加复杂的模型在识别实体上具有更高的精确度[10] [11]。LSTM 与条件随机场(Conditional Random Field, CRF)相结合的模型最近在各个领域的实体识别上受到了很多学者的喜爱[12]。不过目前看来，学者对实体识别技术的研究范围多是针对通用行业领域，仍存在较多特定领域的命名实体识别暂时未得到学术界关注，招标领域就是其中之一。

3. 标的物提取的具体流程

招标领域目前大多依然采用人工的方式处理招标文件文本，消耗了很多不必要的资源，本节主要对实现标的物提取的主要流程进行介绍，主要包括对语料进行的预处理、标的物命名实体的定位、以及基于 CNN 命名实体分类的标的物提取。具体流程图如图 1 所示。

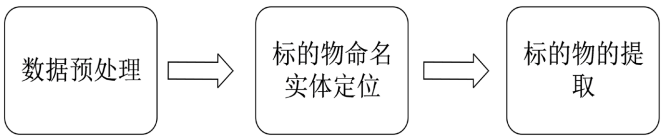

Figure 1. Process of subject matter extraction
图 1. 标的物提取流程

数据预处理：目前，从企业中获得的数据里面包含有各种无序的信息，并且文件的类型也不尽相同，因此首先要对数据进行整理，将文件转化为统一的格式并对其进行整理，同时将一些无用的信息进行清洗。

标的物命名实体定位：招标文件往往包含招标项目的多种信息，如招标条件、项目概况等，标的物仅是包含信息中的一种。若直接对整个招标文件进行命名实体分类，所需要的计算量会很大，且浪费了许多不必要的计算资源，因此需要先对标的物进行初步的定位，从而针对性的对其进行识别提取。

标的物的提取：将含有标的物命名实体的文本片段定位出来后，若只把定位短语作为 CNN 的输入，则会因为上下文信息缺失导致标注结果的准确率较低，因此本文将含有该片段的上下文通过 word2vec 转换为词向量，输入给 CNN，进行 BIOES 标签位置分类，进而再根据位置关系知识，最后提取到核心标的物命名实体。

3.1. 数据预处理

数据预处理阶段的操作主要包括将各种相对无序的信息进行初步整合。本文所用语料主要可以分为两个部分，33,586 份招标文件文本，以及含有 19,980 条招标项目及相关简要信息的分类信息表。为了最终的结构化信息整合阶段更方便快捷的处理数据，本文对语料进行了如下所述预处理。数据的主要处理流程如图 2 所示。

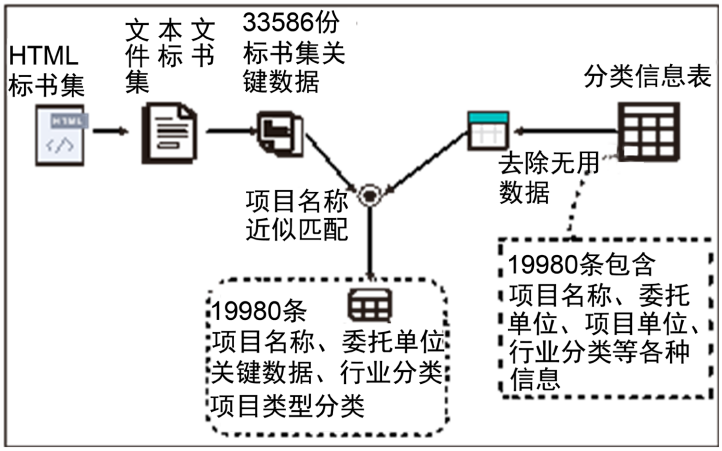

Figure 2. Data processing flow
图 2. 数据处理流程

首先本文所获语料为将招标文件按照日期分别存放的文件夹列表，因此第一步通过批处理将所有文件夹内容进行了合并。其次招标文件语料的初始格式为 HTML (Hyper Text Markup Language)文件，因此本文将 HTML 文件中的招标文件文本部分提取出来并存储到文本文档中以便后续使用，这里使用 python 自带的 HTMLParser 实现，通过定义 HTML 文本内容提取解析器将招标公告文本主体内容解析出来。

因为最终需要输出结构化的信息，所以提取出来的标的物，还要同分类信息表对应起来。将该信息表中项目编号、标段编号等无用信息删除后进行统计分析，发现一方面从招标项目的角度看部分项目缺少分类信息，比如某些项目对应分类结果只有行业分类，另一方面从类别来看，某些类别比如信息化行业类、水电行业类中对应的信息量过少，对模型训练效果的评价不具备参考价值，因此对这部分数据进行了清除。

3.2. 基于正则表达式的标的物命名实体定位

在招标领域，项目所属类型的差异往往也会造成在招标文件中语言使用的差异，因此在进行标的物定位时，思路是先找到货物、服务、工程三个项目类型下招标文件中标的物的相同规则，这样不同项目类型的招标文件主体处理流程也都是相似的。

其流程是先将项目名称中的“#”、“、”、空格、括号及其中内容从中去除，然后通过统计得到每一类中词频最高的词语作为无用数据的分割点，比如货物中是“采购”，而服务类中是“项目”，这类词语及之后的内容大多无用，识别出后裁剪掉。然后将多数语料中都会出现的企业单位名称、项目日期等内容及其前面的部分裁剪掉，最后通过分析中间结果对得到的标的物进行细节上的分析修改，比如货物类中会有“框架”、“框架协议”、“年度”等冗余词，服务类中有“项目(.+)(项目)”等词语结构。具体过程如图 3 所示。

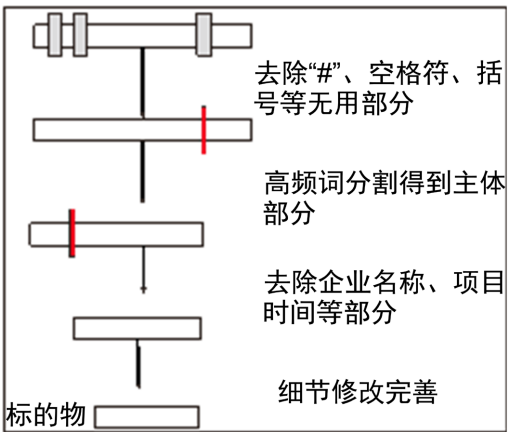

Figure 3. Subject matter located flow
图 3. 标的物定位流程

总的来说，在通过正则匹配方式对标的物进行定位时主要根据表 1 的几个结构进行处理。

Table 1. Regular expression processing structure
表 1. 正则表达式处理结构

正则表达式	作用
/^\s+|\s+$/g	去除空格符
\\(.*?)	去除括号
\\pP|\\pS	去除标点符号
^\d{4}-\d{1,2}-\d{1,2}	时间日期的识别
(项目)(.*?)(项目)	对项目 + 标的物 + 项目的结构进行提取
(集团|公司)(.*?)(项目)	对公司或集团 + 标的物 + 项目的结构进行提取

3.3. 基于知识增强 CNN 的命名实体分类的标的物提取

通过正则表达式对招标文件进行提取之后，会得到一个短语集合。这些短语中有些就是本文所需要的标的物，而有些短语则含有标的物和额外的词汇(如“使用的离子交换树脂采购”)，还有一些短语则并不包含标的物(如“第十二次集中”)。

因为正则表达式是基于规则的，所以无法通过进一步的正则匹配对上述词语进行修改。所以需要使用基于知识增强 CNN 来对短语中的标的物进行分类和提取。基于知识增强的 CNN 命名实体分类模型具体构造图 4 所示。

1) 输入层

经过 2 处理后，将定位到的标的物所在短语与其上下文通过 word2vec 转换为词向量，作为输入。之所以将标的物所在短语的上下文也送入模型，是因为神经网络的特性是能够将文本的上下文语义信息利用起来。比起只把含有标的物的短语送入模型，这样做会让模型的标注结果更加准确。

2) 卷积层

本文通过最大下采样[13]的方法，来捕获卷积层中最明显、最有效的局部特征，从而用来帮助模型实现标的物命名实体识别。因此，对于一个输出为矩阵的卷积层，通过其后的最大下采样层处理之后，得

到的输出结果为：

$$\left[x_\theta^l(\omega)\right]_i = \max_t \left[x_\theta^{l-1}(\omega)\right]_{i,t},\ 1 \le i \le n_{hu}^{l-1} \tag{1}$$

其中 n_{hu}^{l-1} 为卷积层中隐藏单元的个数，即每个窗口所生成的局部特征向量的维度大小。通过下采样层，模型输入维度不统一的问题可以得到解决，并且不需要对待标注词语的上下文信息进行舍弃。

通过最大下采样层得到“最优”局部特征向量之后，该固定维度大小的向量会被输入到标准的神经网络中，并进行最后的标签判别。本文对词句片段进行扩展，在首部和尾部进行扩充，保证每一个词都能“拥有”自己的窗口，从而保证该“边界问题”不会对实验结果产生影响。

3) 实体标注模式

本文根据中文的自然语法对文本中的词进行标注[14]。表 2 列举了标注模式中所有标签的类型以及对应的含义描述。

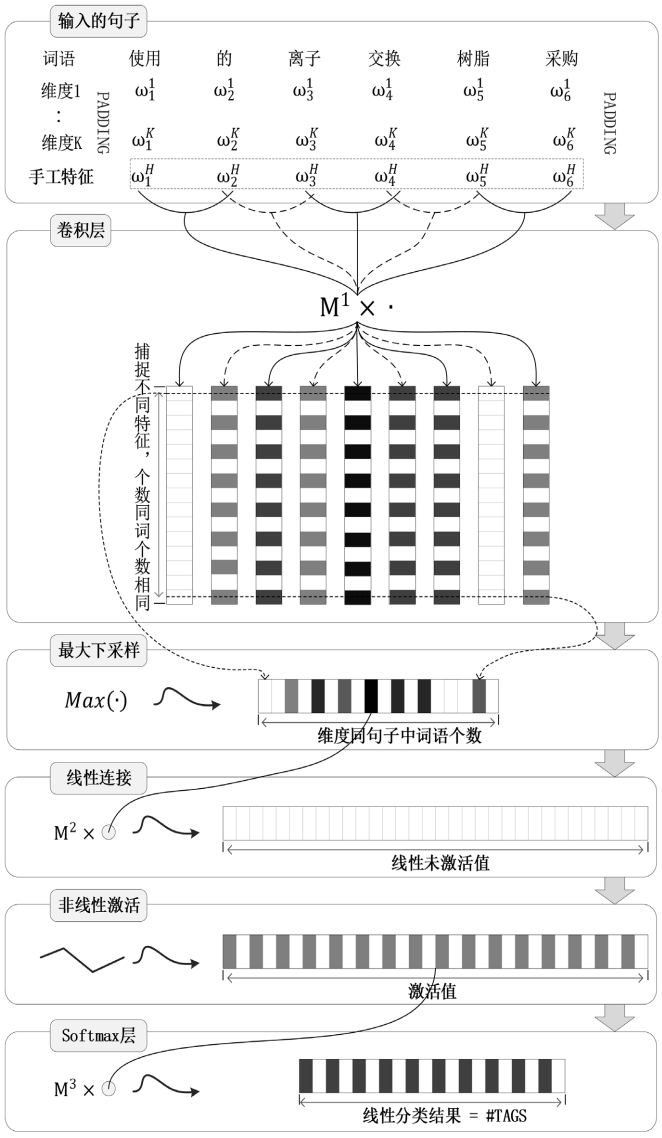

Figure 4. Model architecture
图 4. 模型架构图

Table 2. Type of label and their meaning
表 2. 标签类型及其含义

标签类型	含义
B-begin	表示实体开始
I-inside	表示实体内部
E-end	表示实体尾部
S-single	表示本身就是实体
O-other	表示其他非实体字符

4) Softmax 概率归一化

Softmax 函数的作用便是将线性预测值转换为类别概率，其函数表达式$\sigma(z)=(\sigma_1(z),\sigma_2(z),\cdots,\sigma_k(z))$定义如下：

$$\sigma_i(z)=\frac{\exp(z_i)}{\sum_{j=1}^{k}\exp(z_j)},\ i=1,2,\cdots,k \tag{2}$$

将神经网络的个神经元的输出带入到 Softmax 的结果其实就是对每一个取自然底数幂值，从而变成一个非负的值，然后除以所有项的和进行归一化。最后，Softmax 层的每一个输出都可以看作输入的待标注词语属于标签的概率，或者称为似然(Likelihood)。

5) 获取模型分类结果

通过模型的前述处理之后，包含标的物的定位短语以及其上下文中的每个词语都得到了自己的标签。因为上下文信息只是辅以模型的训练，对实际标的物的提取影响不大，所以本文只针对标的物的定位短语进行分析，以短语“使用的离子交换树脂采购”为例，其经过标注后的数据如表 3 所示：

Table 3. Labeled instance
表 3. 标注实例

短语中的词	标注结果
使用	O-other
的	O-other
离	B-begin
子	I-inside
交	I-inside
换	E-end
树脂	S-single
采购	O-other

表中可以发现，当含有标的物的短语中出现连续的实体时，这些连续的实体就是所需要的标的物。

因此根据这一特点，本文总结了三种不同情况，对短语进行处理：1) 针对全部标注为实体的短语，本文不做修改，把该短语作为所需的标的物；2) 对于在开头或结尾出现不相关词汇的短语，本文将这些不相关词汇进行删除，保留识别出的实体作为标的物；3) 对于无法识别出来实体的短语，本文直接对其进行删除。最终依据以上的流程进行处理后得到符合实际需求的标的物。

4. 实验结果及结果分析

4.1. 实验环境及语料库介绍

本文实验阶段的环境配置是内存为 8G、CPU 为 Intel i7-8565U、显卡为 MX230 的个人电脑。编程语言为 Python3，需要使用不同的第三方库，集成工具使用的是 PyCharm 社区版。

本文实验所用原始资料为：完整招标文件共计有 33,586 份，来自某央企招标网，其中部分招标文件所对应招标项目有分类结果等更详细的信息，部分没有分类结果等信息。有招标项目名称和委托单位等信息、且已知对应的分类结果的项目资料共计 19,980 条，且在完整招标文件集中均有对应招标文件，为保证资料信息数量一致，本文在后期进行了信息整合，实际的使用详情如表 4 所示。

Table 4. Type of label and their meaning
表 4. 标签类型及其含义

行业	数量/个
火电	6538
煤炭	5590
铁路	2368
化工	2632
港口航运	395
新能源	1010
承揽	91
其他	1356

本文研究内容所用全部数据均来自企业，涉及的相关招标项目信息为实际中已经完成的，且分类结果等信息是由企业的相关专业人员根据经验所确定的结果，具有较高的可靠性和可用性。从表 4 中可以看出本文实验各阶段所用语料涵盖不同的行业和不同的项目类型，且各个类别的招标文件具有足够的数量，从而避免了本文所提出的方法的普适性不足，使其能够在后续公司智能化管理中提供一定的指导作用。

4.2. 使用正则表达式对标的物进行初步定位

在使用正则表达式对标的物提取阶段，本文得到的结果如图 5 所示，由于语料库中对于标的物没有确切的信息，因此现阶段只能通过人工方式进行判断，可以看到从大多数招标文件中提取出的标的物是比较准确的，但是由于规则和统计的结果存在一定误差，因此部分招标文件中提取出的标的物不甚准确，比如图 5 中的“第 12 次集中”、“及塔架吊安装工程招标”、“使用的离子交换树脂采购”等。

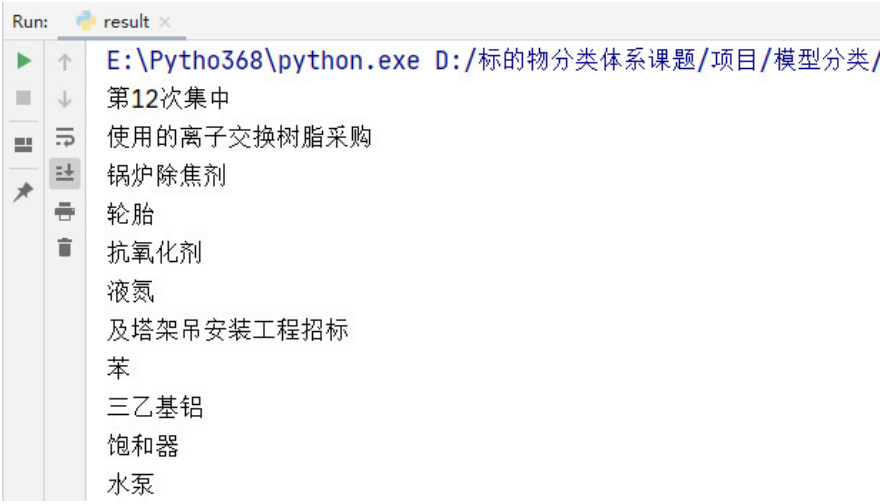

Figure 5. Data processing flow
图 5. 数据处理流程

4.3. 评价指标

本文主要使用精确率、召回率、准确率、F1 分数来评价神经网络对于标的物命名实体分类的效果。在二分类问题中，评价指标的值主要通过 TP (Ture Positive)、FP (False Positive)等计算得到，而在多分类问题中，召回率、F1 分数等的最终结果需要通过宏平均和微平均计算得到，二者在计算方法上有所区别，宏平均是根据每个类计算出召回率、F1 分数然后取平均得到的结果，而微平均是从全部数据中计算出样本被正确或错位预测的数量，类似于将其转化为二分类问题后再计算出评价结果值[15]。本文选择使用宏平均方式进行计算，下面对具体计算方法进行介绍。

以三分类问题为例，其混淆矩阵如表 5 所示。

Table 5. Confusion matrix of three classification problem
表 5. 三分类问题的混淆矩阵

类别	预测为 A 类的个数	预测为 B 类的个数	预测为 C 类的个数
A 类	*AA*	*AB*	*AC*
B 类	*BA*	*BB*	*BC*
C 类	*CA*	*CB*	*CC*

以 A 类为例，通过表 5 混淆矩阵可以计算得到：

$$TP_A = AA \tag{3}$$

$$FP_A = BA + CA \tag{4}$$

$$FN_A = AB + AC \tag{5}$$

其中 TP_A 为 A 类样本被正确预测为 A 类的数量，FP_A 为非 A 类样本被预测为 A 类的数量，FN_A 为 A 类样本被预测为其他类的数量。

$$precision_A = \frac{TP_A}{TP_A + FP_A} \tag{6}$$

$$recall_A = \frac{TP_A}{TP_A + FN_A} \tag{7}$$

$$F1_A = \frac{2 \times precision_A \times recall_A}{precision_A + recall_A} \tag{8}$$

其中 $precision_A$ 为 A 类样本的精确率， $recall_A$ 为 A 类样本的召回率， $F1_A$ 为 A 类样本的 F1 分数。

使用同样的方法可以计算出 B 类的精确率 $precision_B$ 、召回率 $recall_B$ 、F1 分数 $F1_A$ 和 C 类的精确率 $precision_C$ 、召回率 $recall_C$ 和 F1 分数 $F1_C$ ，因此模型最终的评价结果为：

$$precision = \frac{precision_A + precision_B + precision_C}{3} \tag{9}$$

$$recall = \frac{recall_A + recall_B + recall_C}{3} \tag{10}$$

$$F1 = \frac{F1_A + F1_B + F1_C}{3} \tag{11}$$

$$accuracy = \frac{AA + BB + CC}{AA + AB + AC + BA + BB + BC + CA + CB + CC} \tag{12}$$

其中 *precision* 为模型整体的精确率，*recall* 为模型整体的召回率，*F*1 为模型整体的 F1 分数，*accuracy* 为模型整体的准确率。准确率 *accuracy* 可以体现出整体的预测效果，即多少样本被正确预测；精确率 *precision* 体现出对负样本的区分力，值越大区分力越强；召回率 *recall* 体现出对正样本的区分力，值越大区分力越强；F1 分数则是对精确率 *precision* 和召回率 *recall* 的整体评价。

4.4. 基于知史增强 CNN 的命名实体分类的标的物提取

如表 6 所示，分别使用不同的神经网络对实体命名分类的评价结果。

Table 6. Comparison of different neural network algorithms
表 6. 不同神经网络算法对比

神经网络	精确率/%	召回率/%	准确率/%	F1 分数/%
DNN	91.42	91.22	91.59	91.32
RNN	95.39	95.43	95.43	95.41
HNN	95.40	95.22	95.49	95.31
CNN	95.19	95.65	95.81	95.42

可以看到在基于深度神经网络(Deep Neural Networks, DNN)、循环神经网络(Recurrent Neural Networks, RNN)、霍普菲尔得神经网络(Hopfield Neural Network, HNN)、CNN 的神经网络在实体命名分类上 CNN 表现最好，其准确率相比其三个网络中最高的 HNN 还要高出 0.3%。本文同时还对不同行业以及不同项目类型的数据进行了十折交叉验证，得到的准确率均在 95.8%左右，说明本方案具有较好的鲁棒性。

4.5. 结果评价

为了评价方法效果，从 19,980 条招标文件文本中随机抽取 1% (即 200 份)招标文件，并将这些招标文件进行效果评估，得到的结构化字段的准确率和召回率均高于 90%，且货物类招标文件相对更为规范，

该类招标文件的准确率和召回率更高。采用基于规则的方法对招标文件文本文件进行信息抽取较为适宜，最终实现标的物提取。

5. 结束语

本文在实现标的物提取时，为了使提取效果不受招标文件所涉及行业领域宽泛的影响，采用了传统的基于规则办法实现标的物提取，主要通过正则匹配的方式实现，并采用基于知识增强 CNN 的命名实体识别技术对定位的标的物短语进行修改，通过将该方法所提取的标的物与人工标注的标的物进行比对，发现该方法能够在标的物提取时取得良好效果。在未来的工作中，将尝试对语料进行一步研究，在神经网络的训练过程中，加入人工特征，进一步提高在实际场景中的准确率。

参考文献

[1] Zhang, Y. and Yang, J. (2018) Chinese NER Using Lattice LSTM. In: *Proceeding of 56th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, 1554-1564. https://doi.org/10.18653/v1/P18-1144

[2] Strubell, E., Verga, P., Belanger, D. and Mccallum, A. (2017) Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. https://arxiv.org/pdf/1702.02098.pdf

[3] Zhu, Y. and Wang, G. (2019) CAN-NER: Convolutional Attention Network for Chinese Named Entity Recognition. In: *Proceedings of the* 2019 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Association for Computational Linguistics, Stroudsburg, 3384-3393.

[4] Cao, P., Chen, Y., Liu, K., Zhao, J. and Liu, S. (2018) Adversarial Transfer Learning for Chinese Named Entity Recognition with Self-Attention Mechanism. In: *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, 182-192. https://doi.org/10.18653/v1/D18-1017

[5] Peng, N. and Dredze, M. (2017) Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning. https://arxiv.org/pdf/1603.00786v2.pdf

[6] Fei, R., Guo, J., Wang, C. and Sun, Y. (2020) Research on Chinese Electronic Medical Record Named Entity Recognition Based on Lexicon Enhancement. *International Journal of Education and Teaching Research*, **1**, 176-182.

[7] Xu, H., Liu, H., Yang, G. and Zhang, C. (2017) Sentiment Analysis of Chinese Version Using SVM & RNN. In: *Proceedings of the* 6*th International Conference on Information Engineering* (*ICIE* '17). ACM, New York, 1-5. https://doi.org/10.1145/3078564.3078565

[8] 陈曦. 基于文本信息抽取的高铁车载设备故障发现的理论与方法[D]: [硕士学位论文]. 北京: 北京交通大学, 2017: 15-19.

[9] 祖木然提古丽·库尔班. 基于神经网络的电子病历实体识别[D]: [硕士学位论文]. 乌鲁木齐: 新疆大学, 2019: 2-3.

[10] Ratinoy, L. and Roth, D. (2009) Design Challenges and Misconceptions in Named Entity Recognition. In: *Proceedings of the* 3*th Conference on Computational Natural Language Learning*, ACM, New York, 147-155. https://doi.org/10.3115/1596374.1596399

[11] Dai, Z., Yang, Z., Yang, Y., *et al*. (2019) Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, 28 July-2 August 2019, 2978-2988. https://doi.org/10.18653/v1/P19-1285

[12] Didrik, N. (2016) Tree Boosting with XGBoost-Why Does XGBoost Win "Every" Machine Learning Competition? Norwegian University of Science and Technology, Trondheim.

[13] 万小军, 冯岩松, 孙薇薇. 文本自动生成研究进展与趋势[C]//CCF2014-2015 中国计算机科学技术发展报告会论文集. 北京: 机械工业出版社, 2015: 298-323.

[14] 郁亚辉. 产品评论挖掘中特征同义词的识别[J]. 中文信息学报, 2016, 30(4): 150-158.

[15] Saha, S. and Ekbal, A. (2013) Combining Multiple Classifiers Using Vote Based Classifier Ensemble Technique for Named Entity Recognition. *Data & Knowledge Engineering*, **85**, 15-39. https://doi.org/10.1016/j.datak.2012.06.003